DOCUMENT RESUME

ED 337 464 TM 017 261

AUTHOR Hanson, Bradley A.
TITLE An Investigation of Methods for Improving Estimation of Test Score Distributions.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa.
REPORT NO ACT-RR-90-4
PUB DATE Jun 90
NOTE 32p.
AVAILABLE FROM ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Simulation; *Equations (Mathematics); *Estimation (Mathematics); Licensing Examinations (Professions); *Mathematical Models; Maximum Likelihood Statistics; Multiple Choice Tests; Population Distribution; *Scores; *Statistical Distributions
IDENTIFIERS *Beta Binomial Test Model; Kernel Method; *Observed Frequencies (Test); Smoothing Methods

ABSTRACT

Three methods of estimating test score distributions that may improve on using the observed frequencies (OBFs) as estimates of a population test score distribution are considered: the kernel method (KM); the polynomial method (PM); and the four-parameter beta binomial method (FPBBM). The assumption each method makes about the smoothness of the true distribution and computational details of the methods are described. The methods are compared in a simulation study in which 500 samples of size 500, 1,000, 2,000, and 5,000 are taken from three population distributions. For each of the 6,000 samples (3 population distributions by 4 sample sizes by 500 samples), the variable kernel, fixed kernel, FPBBM, and PM estimates of the true distribution are computed. The three population distributions are defined using observed raw score distributions on three tests for which a large number of examinees is available. Methods based on smoothness assumptions performed better than using the OBFs, and differences among the methods' performance were small compared to differences between the performance of the worst method and use of OBFs. The FPBBM performed best across all conditions, although the PM performed as well as the FPBBM for sample sizes of 5,000. The PM generally performed better than the KM except for one population for which the test score distribution was relatively flat. Four tables, 15 graphs, and a 13-item list of references are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ACT Research Report Series 90-4

ED337464

An Investigation of Methods for Improving Estimation of Test Score Distributions

Bradley A. Hanson

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. FERGUSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

June 1990

TM017261

BEST COPY AVAILABLE

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

© 1990 by The American College Testing Program. All rights reserved.

AN INVESTIGATION OF METHODS FOR IMPROVING ESTIMATION OF TEST SCORE DISTRIBUTIONS

Bradley A. Hanson

ABSTRACT

This paper considers three methods of estimating test score distributions that potentially improve upon the observed frequencies as estimates of a population test score distribution: the kernel method, the polynomial method, and the 4-parameter beta binomial method. The assumption each method makes about the smoothness of the true distribution and computational details of the methods are described. The methods are compared with a simulation study in which 500 samples of size 500, 1000, 2000, and 5000 are taken from each of 3 population distributions. The three population distributions are defined using observed raw score distributions on three tests for which a large number of examinees are available. All the methods based on smoothness assumptions performed far better than using the observed frequencies. The differences among the performance of the methods were small compared to the difference between performance of the worse performing method and using observed frequencies. The 4-parameter beta binomial method performed best in the simulation study across all conditions, although the polynomial method performed equivalently for sample sizes of 5000. The polynomial method generally performed better than the kernel method except for one of the populations for which the test score distribution was relatively flat. Conclusions and suggestions are offered concerning the use of the methods in practice.

This paper will investigate three methods for potentially improving on observed frequencies as estimates of a population raw test score distribution. These methods have the potential to provide better results than using the observed frequencies in applications in which the estimates of the population raw test score distribution are used. Examples of such applications are: describing and comparing raw test score distributions, constructing norms, and equating using equipercentile methods (Kolen, 1988).

The potential improvement offered by each of the three estimation methods is based on introducing assumptions about the smoothness of the population distribution. The methods are distinguished by the specific assumptions about the smoothness of the population distribution that are made. The success of each method in any specific case will depend on the appropriateness of the smoothness assumptions made by the method for that case.

In order to evaluate the improvement of an estimation technique based on smoothness assumptions over using observed frequencies it is necessary to have a criterion to measuring the performance of an estimation method. In this paper the primary measure of the performance of an estimation method for a particular sample will be average squared error defined as

$$ASE = \frac{1}{K+1} \sum_{i=0}^{k} (\tilde{f}(i) - f(i))^2 , \qquad (1)$$

where K is the number of items on the test, f(i) is the true probability of raw score i, and $\tilde{f}(i)$ is an estimate of f(i). Across samples, the expected value of ASE (E[ASE]) will be used as a criterion. E[ASE] will be referred to as mean squared error (MSE). The mean squared error can be written as the sum of two quantities to be referred to as bias squared and variance

$$MSE = \frac{1}{K+1} \sum_{i=0}^{k} [f(i) - E(\tilde{f}(i))]^2 + \frac{1}{K+1} \sum_{i=0}^{k} E[\tilde{f}(i) - E(\tilde{f}(i))]^2$$

$$= \text{bias}^2 + \text{variance} . \qquad (2)$$

Methods of discrete density estimation based on smoothness assumptions have reduced variance as compared to the observed frequencies but generally introduce bias (whereas the observed frequencies are unbiased estimates of the true frequencies).

The next section describes three density estimation methods based on smoothness assumptions. The following section describes a simulation study comparing the three methods.

Density Estimation Methods

This section describes the smoothness assumptions that are the basis of the three density estimation methods to be compared and the computational details for each method.

Kernel Techniques

If $\hat{f}(i)$ is the probability of raw score, i, i = 0, ..., K, in the observed sample, then the kernel estimate of the true probability of raw score i (f(i)) is given by a local weighted average of $\hat{f}(j)$, for j in a neighborhood of i. Specifically, the locally weighted average of $\hat{f}(j)$ is computed as

$$a(i) = \sum_{j=i-h_i/2}^{i+h_i/2} w_i(j - i + \frac{h_i}{2}) \hat{f}(j) , \qquad (3)$$

where $\hat{f}(j) \equiv 0$ if $j < 0$ or $j > K$, and h_i (an even positive integer) is a parameter that determines the width of the local neighborhood of raw score i over which the weighted average is taken. The $w_i(k) > 0$, k = 0, ..., h, are referred to as the kernel for raw score i. The $w_i(k)$ are taken to sum to 1 over k, so that the kernel is a discrete probability distribution. The kernel

estimate of f(i) is given in terms of the a(i) as

$$\tilde{f}_k(i) = \frac{a(i)}{\sum_{j=0}^{k} a(j)} \quad . \tag{4}$$

If $\hat{f}(i) = f(i) + e(i)$, where $e(i) \equiv \hat{f}(i) - f(i)$, then the smoothness assumption under which the kernel technique works well is that the variation in f(i) in the neighborhood of i used to compute a(i) is "small" compared to the variation in e(i) in that neighborhood. To see this write $\tilde{f}_k(i)$ as

$$\tilde{f}_k(i) = \sum_{j=i-h_i/2}^{i+h_i/2} w_i^*(j - i + (h_i/2))[f(j) + e(j)] \ ,$$

$$\text{where } w_i^*(j - i + (h_i/2)) = \frac{w_i(j - i + (h_i/2))}{\sum_{j=0}^{k} a(j)} \quad . \tag{5}$$

From Equation 5 it is seen that the variance of $\tilde{f}_k(i)$ only depends on the terms $w_i^*(j - i + (h_i/2))\ e(j)$. Hence, the variance of $\tilde{f}_k(i)$ will, in most practical situations, be less than the variance of $\hat{f}_k(i)$ due to $w_i^*(j - i + (h_i/2))$ being less than 1 for all j. The bias of $\tilde{f}_k(i)$ is given by the difference of f(i) and the sum of $w_i^*(j - i + (h_i/2))\ f(j)$, which will not in general be zero. The kernel technique will work well for estimating f(i) when the bias introduced is less than the reduction of variance, which will occur when f(j) is smooth in the sense that for the sample size under consideration the variation of f(j) is less than the variation in e(j) in the neighborhood of i used to compute $\tilde{f}_k(i)$.

Application of the kernel method to produce a density estimate requires selection of the kernel ($w_i(k)$) and window widths (h_i) for all raw score points i. In this paper the binomial kernel will be used. For window width

h, the binomial kernel is $w(k) = \text{prob}(Z = k)$, where the random variable Z has a binomial distribution with parameters h and .5.

Given the kernel to be used, producing a kernel density estimate then reduces to selecting the window widths (h_i) to use at each raw score point. This task is made simpler by requiring $h_i = h$ be constant for all raw score points. The kernel method with h constant across raw score points will be referred to as the fixed kernel method. The kernel method in which the h_i are allowed to vary across raw score points will be referred to as the variable kernel method.

The strategy to be employed in choosing h for the fixed kernel method will be to estimate ASE for different values of h and choose the h such that the estimated ASE is minimized. Cross validation will be used to estimate ASE.

Efron (1983, remark B) and Wong (1983) discuss using bootstrap, jackknife and cross validation methods for choosing smoothing parameters in density estimation. The discussion here will follow Efron (1983). Writing $(\tilde{f}(i) - f(i))^2 = \tilde{f}(i)^2 + f(i)^2 - 2\,\tilde{f}(i)\,f(i)$, it can be seen that minimizing ASE is equivalent to minimizing the quantity

$$\text{Err} = \sum_{i=0}^{k} [\tilde{f}(i)]^2 - 2 \sum_{i=0}^{k} f(i)\,\tilde{f}(i) \quad . \tag{6}$$

An obvious estimate of Err is given by replacing the unknown quantity f(i) with $\hat{f}(i)$. This estimate will tend to underestimate Err since $\tilde{f}(i)$ is computed using $\hat{f}(i)$, and will be referred to as the over-optimistic estimate of Err (and denoted err). For the purposes of choosing a window width, using err as an estimate of Err will always result in the smallest window width being chosen. Let $\omega \equiv E(\text{Err} - \text{err})$ be the expected amount by which err

underestimates Err. If ω were known then an estimate of Err, which would be better than using err, could be obtained as: err + ω.

The bootstrap estimate of ω is obtained by taking the expectation used to define ω over $\hat{f}(i)$ rather $f(i)$. The bootstrap estimate of ω is given by

$$\hat{\omega}^{\text{boot}} = 2\, E^{*}\left[\sum_{i=0}^{k} \tilde{f}^{*}(i)\,(\hat{f}^{*}(i) - \hat{f}(i))\right], \tag{7}$$

where the "*" indicates sampling from $\hat{f}(i)$. Therefore, $\tilde{f}^{*}(i)$ is an estimate of the true density based on a sample density $\hat{f}^{*}(i)$, which is a sample from $\hat{f}(i)$, and E^{*} indicates an expectation over sampling from $\hat{f}(i)$. $\hat{\omega}^{\text{boot}}$ can be computed by Monte Carlo methods as follows. B samples $(\hat{f}^{*b}(i),\ b = 1, \ldots, b)$ are simulated from $\hat{f}(i)$ and the estimated density $(\tilde{f}^{*b}(i))$ is computed for each sample. $\hat{\omega}^{\text{boot}}$ is then approximated by

$$\frac{2}{B}\sum_{b=1}^{B}\sum_{i=0}^{K} \tilde{f}^{*b}(i)\,(\hat{f}^{*b}(i) - \hat{f}(i)) \quad . \tag{8}$$

$\hat{\omega}^{\text{boot}}$ is added to err to produce an estimate of Err.

The jackknife estimate of ω is a second order approximation of $\hat{\omega}^{\text{boot}}$ (Efron, 1983, 1982) given by

$$\hat{\omega}^{\text{jack}} = \frac{2}{K+1}\sum_{j=0}^{K}\sum_{i=0}^{K}\left[\tilde{f}_{-j}(i)\,\hat{f}_{-j}(i) - \tilde{f}_{-j}(i)\,\hat{f}(i)\right]$$

$$\text{where } \hat{f}_{-j}(i) = \begin{cases} \frac{n}{n-1}\,\hat{f}(i) - \frac{1}{n-1} & j=i \\ \frac{n}{n-1}\,\hat{f}(i) & j\neq i \end{cases}, \tag{9}$$

n is the sample size, and $\tilde{f}_{-j}(i)$ is the density estimate using $\hat{f}_{-j}(\ell)$ in place of $\hat{f}(\ell)$, $\ell = 0, \ldots, K$. The jackknife estimate of ω is less computationally burdensome than computing the bootstrap estimate by Monte Carlo methods.

The cross-validation estimate of Err is a modification of err with $\tilde{f}_{-i}(i)$ in place of $\tilde{f}(i)$ in the expression involving $\hat{f}(i)$:

$$\hat{Err}^{cv} = \sum_{i=0}^{K} [\tilde{f}(i)]^2 - 2 \sum_{i=0}^{K} \hat{f}(i)\, \tilde{f}_{-i}(i) \tag{10}$$

This modification should reduce the overestimation of err. The cross-validation estimate of ω is given by:

$$\tilde{\omega}^{cv} = 2 \sum_{i=0}^{K} [\tilde{f}(i)\, \hat{f}(i) - \tilde{f}_{-i}(i)\, \hat{f}(i)] \tag{11}$$

Comparing Equations 11 and 9 it is expected that the cross-validation estimate of Err and the jackknife estimate of Err will be similar. Efron (1983) found this to be true with cross-validation and jackknife estimates of prediction error in a dichotomous prediction problem. As is seen by comparing Equations 11 and 9, the cross-validation estimate of Err requires less computation than the jackknife estimate.

Kolen (1988) shows that for the fixed kernel density estimate a simple approximate expression exists for the cross-validation estimate of Err which only requires one kernel estimate be computed for each value of h. The expression derived here uses a slightly different approximation than the expression given by Kolen (1988), but the two expressions seem to produce very similar results.

First, note that if $\tilde{f}_{k_h}(i)$ is the fixed kernel estimate with window width h then $\tilde{f}_{k_h,-i}(i)$ (the estimate based on $\hat{f}_{-i}(\ell)$ rather than $\hat{f}(\ell)$, $\ell = 0, \ldots, K$) can be written as

$$\tilde{f}_{k_h,-i}(i) = \frac{1}{\sum_{j=0}^{K} a_{-j}(j)} \sum_{j=i-h/2}^{i+h/2} w(j - i + (h/2))\, \hat{f}_{-i}(j)$$

$$= \frac{1}{\sum_{j=0}^{K} a_{-j}(j)} \sum_{j=i-h/2}^{i+h/2} w(j - i + (h/2)) \left[\frac{n}{n-1}\hat{f}(j) - \frac{w(h/2)}{n-1}\right], \qquad (12)$$

where $a_{-i}(i)$ is given by Equation 3 with $\hat{f}_{-i}(i)$ substituted for $\hat{f}(i)$. If the sum of $a_{-j}(j)$ is approximately equal to the sum of $a(j)$ then $\tilde{f}_{k_h,-i}(i)$ can be approximated by

$$\tilde{f}_{k_h,-i}(i) \approx \frac{n}{n-1}\tilde{f}_{k_h}(i) - \frac{w^*(h/2)}{n-1}. \qquad (13)$$

Substituting this approximate expression for $\tilde{f}_{-i}(i)$ in Equation 10 gives an approximate cross-validation estimate of Err for the kernel estimate with window width h:

$$\hat{Err}_h^{cv} \approx \sum_{i=0}^{K} [\tilde{f}_{k_h}(i)]^2 - \frac{2n}{n-1}\left[\sum_{i=0}^{K} \hat{f}(i)\tilde{f}_{k_h}(i) - \frac{w^*(h/2)}{n}\right]. \qquad (14)$$

The expression for the approximate cross-validation estimate of Err given by Equation 14 is computationally efficient in that it only requires the kernel estimate to be computed once for each h.

The fixed kernel estimate used in this paper is obtained by computing the kernel estimate (using the binomial kernel) for h = 2, 4, ..., 36 and choosing the h that minimizes Equation 14.

Table 1 presents results from a small simulation study comparing the performance of Equation 14 with the bootstrap estimate of Err using $\hat{\omega}^{boot}$ from Equation 7 in choosing the value of h that minimizes Err for the kernel estimate (200 bootstrap samples were used). Ten samples of size 1000 were simulated from a 4-parameter beta binomial distribution (Lord, 1965) that was fit to a raw score distribution of 980 college-bound 10th grade examinees on a recent form of the 50-item P-ACT+ writing test. For each of the 10 samples, fixed kernel estimates using values of h from 2 to 36 were computed and Err was estimated with the bootstrap using Equation 7 and cross validation using Equation 14. Table 1 presents the values of h and corresponding values of Err chosen by both the bootstrap and cross validation along with the h that produces the minimum Err for each sample. For 6 of the 10 samples the kernel estimate using the h picked by the bootstrap has lower Err than the corresponding estimate using the h picked by cross validation. For two of the samples the estimate using the h picked by cross validation has lower Err than the estimate using the h picked by the bootstrap. For the remaining two samples the same value of h is picked by both cross validation and the bootstrap. The average error across samples using cross validation is slightly lower than the average error using the bootstrap.

Even though the results reported in Table 1 are very limited, they suggest that choosing h using the approximate cross-validation formula (14) can work fairly well compared to the bootstrap, with sample sizes of around 1000.

Estimates of Err can also be obtained at each score point. Equating 14 can be written as

$$\hat{Err}_h^{cv} = \sum_{i=0}^{K} \hat{Err}_h^{cv}(i) \quad , \tag{15}$$

12 A

where $\hat{Err}_h^{cv}(i) \equiv [\tilde{f}_{k_h}(i)]^2 - 2\,\hat{f}(i)\left[\frac{n}{n-1}\,\tilde{f}_{k_h}(i) - \frac{1}{n-1}\,w^*(h/2)\right]$.

Each $\hat{Err}_h^{cv}(i)$ is a cross-validation estimate of $\hat{Err}_h(i) \equiv [\tilde{f}_{k_h}(i)]^2 - 2f(i)\,\tilde{f}_{k_h}(i)$. Equation 15 could be used in a manner analogous to the use of Equation 14 in the fixed-kernel method to choose a value of h for each raw score value.

The variable kernel estimate used in this paper is obtained by computing the kernel estimate for values of h from 2 through 36 and at each raw score point i, choosing the h_i that minimizes $\hat{Err}_h^{cv}(i)$. The resulting values of h_i are then smoothed as a function of i using a robust 3RSS median smoother twice (Tukey, 1977). The purpose of smoothing the values of h_i chosen by cross validation is to reduce large fluctuations in the values of h_i as a function of i.

Polynomial Method

The polynomial smoothing method is based on the following smoothness assumption:

$$\log(f(i)) = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \ldots + \alpha_d i^d \quad , \tag{16}$$

where d is small relative to K. The true density is assumed to be smooth in the sense that its log is a low order polynomial function of the raw score. When d = K Equation 16 will hold for any f(i); i.e., in this case Equation 16 is a representation of f(i) rather than a model. Haberman (1974) discusses estimation and the selection of d for a generalization of the model of Equation 16 in which two ordered categories are modeled. Rosenbaum and Thayer (1987) discuss using the Haberman model for the estimation of bivariate raw test score distributions.

There are two nice properties of maximum likelihood estimates of the parameters in Equation 16 (Darroch and Ratcliff, 1972). First, if maximum

likelihood is used to fit a model of degree d then the first d moments of the estimates and observed distributions will be the same. Second, the maximum likelihood estimate of the distribution given by Equation 16 is the distribution with maximum entropy of those discrete distributions whose first d moments are equal to the first d moments of the observed distribution.

The polynomial method of density estimation using the distribution given by Equation 16 involves two steps: selection of a value of d, and estimation of the parameters in Equation 16 by maximum likelihood using the value of d chosen.

In this paper, Haberman's (1974) model selection strategy will be used to choose d. First, the value of q is chosen such that the true density fits the model given by Equation 16 with d at most equal to q. If L_i^2 is the likelihood ratio chi-square for the maximum likelihood fit of the model of degree i then for j = 2, ..., q, $L_{j-1}^2 - L_j^2$ is the likelihood ratio chi-square for the null hypothesis H_{j-1} versus H_j, where H_j is the hypothesis that the model of degree j is true. Haberman (1974) states that if H_{j^*} is true then the statistics $L_{j-1}^2 - L_j^2$ for $j = q, q - 1, \ldots, j^* + 1$ are asymptotically independent chi-square distributions with 1 degree of freedom. For a level of significance γ, with $\gamma^* = 1 - (1 - \gamma)^{1/(q-1)}$, the probability that $L_{j-1}^2 - L_j^2$, $j = q, q - 1, \ldots, j^* + 1$ exceeds C, the upper γ^* percentage point for the chi-square distribution with 1 degree of freedom is asymptotically no greater than γ. A simultaneous test of the hypotheses H_j, $j = q - 1, q - 2, \ldots, 1$, is to reject all hypotheses H_j such that $j < j'$, where j' is the largest j such that $L_{j-1}^2 - L_j^2 > C$. With values of q and γ specified, this hypothesis testing procedure would allow one to eliminate from consideration models with degrees less than j'; it gives no guidance for choosing from among

the models with degree greater than or equal to j' and less than or equal to q.

The rationale of Haberman's model selection procedure, based on identifying the "true" model, is not necessarily consistent with the goal of selecting a model that minimizes the average squared error (Equation 1) for a sample. Even if there were a model with degree less than K that produced an estimate that was identical to the true density, there is no reason why this degree would minimize the average squared error for a particular sample, a degree higher or lower than the true degree may have a smaller average squared error for the sample. Using a bootstrap, jackknife, or cross-validation estimate of Err to choose a model might work better than the hypothesis testing procedure, although any of these resampling procedures would require extensive computation. There is no formula analogous to that of Equation 14 for the fixed kernel method that would give an estimate of Err based on fitting model given by Equation 16 one time for each degree.

In this paper values of q = 10 and γ = .10 will be used for the polynomial estimation method. The model with the smallest number of parameters that is not rejected by the Haberman procedure will be used as the degree of polynomial to be fit. Equation 16 will be estimated by maximum likelihood using the Newton-Raphson algorithm (Haberman, 1974) for the degree of polynomial chosen.

Beta Binomial Method

The 4-parameter beta binomial method makes the most specific smoothness assumptions of the methods considered in this paper. This method assumes the true density is a member of a four parameter family of smooth densities: the four parameter beta binomial distribution (Lord, 1965). The rationale behind the use of this family of densities is a strong true score model in which

examinee true scores are assumed to have a four parameter beta distribution, and the distribution of raw scores conditional on true score is binomial. A two-term approximation to the compound binomial can be used in place of the binomial, but in practice this does not seem to improve the fit to raw score distributions over using the binomial (Lord, 1965). Two of the parameters of the four parameter beta distribution (denoted p and q) determine the shape of the distribution (i.e., they completely determine all properties of the distribution except scale and location). The remaining two parameters (denoted a and b) are the lower and upper limits of the proportion correct true score distribution. The true score distribution has nonzero density only between the lower and upper limits.

In this paper, the method of moments will be used to estimate the four parameters (Lord, 1965). First, the formula for the relation of raw score to true score moments given by Lord (1965) is used to produce estimates of the first four true score moments from estimates of the first four raw score moments. The parameters are then calculated using an expression giving the parameters in terms of the first four true score moments. This expression is obtained by solving the equations giving the mean, variance, skewness, and kurtosis of a 4-parameter beta distribution in terms of p, q, a, and b for the parameters. The solution for p and q given by Johnson and Kotz (1970, their Equation 13, page 41) is incorrect. The correct solution for p and q can be obtained using their expression for $r = p + q$ (which is correct) given above their Equation 13 to solve their Equations 8.3 and 8.4 for p and q. The expressions for parameters a and b are then obtained using these expressions for p and q and solving the equations giving the mean and variance in terms of p, q, a, and b for a and b.

In some cases, the expression for the parameters in terms of the true score moments will not have a solution, or the solution will contain one or more invalid parameter values (e.g., an upper limit greater than 1). In this case, parameter estimates are found by requiring that the mean, variance, and skewness of the observed and fitted distribution agree, which will determine three of the four parameters (p, q, and b are used here), and choosing the value of the fourth parameter (a) that minimizes the squared difference in the observed and fitted kurtosis. This procedure was successful in producing parameter estimates for all the actual and simulated data sets considered in this paper, i.e., in all cases parameter estimates could be found such that at least the first three moments of the observed and fitted raw score distributions agreed.

Equation (59) of Lord (1964) was used to compute the estimated 4-parameter beta binomial raw score distribution using the 4 estimates parameters of the beta distribution (this expression for the raw score distribution does not appear in Lord, 1965).

Illustration of the Estimation Methods

Figures 1 and 2 present 4-parameter beta binomial, polynomial, fixed kernel, and variable kernel estimates for two data sets. Figure 1 presents the raw score distribution for 3039 examinees for a recent form of the ACT Mathematics test. Figure 2 presents the raw score distribution for 1727 11th grade examinees on a recent form of the P-ACT+ Writing test (this data was obtained from a special study in which the P-ACT+ was administered). For the data in Figure 1 a polynomial of degree 5 and a fixed kernel window width of 8 were chosen using the procedures described above. For the data in Figure 2 a polynomial of degree 6 and a fixed kernel window width of 2 were chosen.

For both data sets the 4-parameter beta binomial and polynomial methods produced good fits. The kernel methods provided good fits for the data in Figure 1, although the kernel fits were more bumpy than the fits of the other two methods. The window widths chosen by the kernel methods for the data in Figure 2 do not seem to provide enough smoothing.

A Study Comparing the Estimation Methods

The primary criterion used in this paper for evaluating the performance of an estimation method across samples is the mean squared error given in Equation 2. A simulation study similar to that of Cope and Kolen (1987) is used here to provide some information on the relative performance of the estimation methods in terms of the mean squared error for some realistic situations. Observed raw score distributions for tests for which data from a very large number of examinees are available are used as population distributions. Monte Carlo methods are used to estimate the mean squared errors for each of the methods for several samples sizes.

Data

Data from three tests will be used as population distributions. The first test is a 200-item multiple choice licensure test. Due to the amount of computation involved in computing the estimates for a large number of samples for a 200-item test, only the 59 internal anchor items will be used. The 59-item internal anchor is designed as a shorter parallel version of the full 200-item test. For this study responses to 39,149 examinees from a recent test date were used as a population distribution.

For the other two population distributions, data from the responses of 230,065 examinees on the Mathematics and Social Sciences tests for a recent October administration of the ACT Assessment were used. Population distributions for each test were defined as the observed frequency

distributions of raw scores. Examinees with zero raw scores (5 examinees for the Mathematics test and 33 examinees for the Social Sciences test) were not used in order to eliminate examinees who did not respond to any of the items on the test or whose responses were hand scored, in which case a raw score of zero was reported in the data set used. This had the effect of setting f(0) equal to 0 for both tests.

Figure 3 presents the three population distributions used in this study.

Method

For each of the three population distributions 500 samples with sample sizes 500, 1000, 2000, and 5000 were simulated. For each of the 6000 samples (3 population distributions by 4 sample sizes by 500 samples) the variable kernel, fixed kernel, 4-parameter beta binomial and polynomial estimates of the true distribution were computed. Thus, including the observed frequencies, five estimates of the population distribution were computed for each sample.

For each estimation method, sample size, and population distribution the average of the values of ASE over the 500 samples was taken as an estimate of MSE. In addition, bias squared and variance were also estimated as averages over the 500 samples of the appropriate values given in Equation 2.

The maximum absolute difference between the population and estimated cumulative relative frequencies was also computed for each estimation method in each sample. The average of these values over the 500 samples will be used as an additional criterion in judging the relative performance of the estimation methods (this criterion will be denoted K-S because it is based on a Komolgorov-Smirnov type statistic (Conover, 1980)). The reason for considering K-S is that for some applications, such as calculating norms and

equipercentile equating, the estimated cumulative relative frequencies rather than the relative frequencies are used.

Results

Tables 2, 3, and 4 contain the results for the populations corresponding to the licensure test, ACT Mathematics test, and ACT Social Science test, respectively. In each table estimates computed using the 500 samples for each sample size of MSE, bias squared, variance and K-S are reported for each of the five estimation methods (in addition, the standard errors of K-S are given). Figures 4 and 5 contain plots of estimates of MSE_i and $variance_i$ as a function of raw score for the licensure and ACT Mathematics tests, respectively, for a sample size of 1000, where

$$MSE_i = E[\tilde{f}(i) - f(i)]^2$$

$$variance_i = E[\tilde{f}(i) - E(\tilde{f}(i))]^2 \ . \qquad (17)$$

In the following discussion of the results, MSE and K-S are not distinguished because of the similarity of the results for the two criteria.

Tables 2 through 4 show that all the estimation methods based on smoothness assumptions performed better than using the observed frequencies. For samples sizes less than 5000 the 4-parameter beta binomial method performed better than the other methods. For a sample size of 5000 the 4-parameter beta binomial and polynomial methods performed about equivalently and better than the kernel methods.

The polynomial method had the lowest bias of the four methods based on smoothness assumptions for all cases in Tables 2, 3, and 4, except for sample sizes of 500 and 1000 for the ACT Mathematics test. The 4-parameter beta binomial method had the lowest variance of all the methods for all cases in Tables 2, 3, and 4.

The variable kernel method did not consistently perform better than the fixed kernel method. For the licensure test the fixed kernel method performed better than the variable kernel method for all sample sizes. For the ACT tests the variable kernel method tended to have lower MSE than the fixed kernel method for the higher sample sizes (2000 and 5000 for mathematics and 1000, 2000, and 5000 for social sciences), and lower bias but greater variance than the fixed kernel method for sample sizes of 2000 and less.

The polynomial method performed worse than the kernel methods for the ACT Mathematics test for all sample sizes except 5000 and performed worse than the fixed kernel method for the ACT Social Science test for the sample size of 500. This is in contrast to the polynomial method performing better than the kernel methods in all other cases.

The results in Table 3 for the ACT Mathematics test can be compared to the results in Table 3 of Cope and Kolen (1987), which reports results analogous to those reported here from a simulation study using as a population distribution data from an October administration (from a different year than used here) of the ACT Mathematics test. The values MSE and K-S in Cope and Kolen's Table 3 for the observed relative frequencies and the 4-parameter beta binomial method (the only two estimation methods in common with the present study) are very similar to those reported here in Table 3. Cope and Kolen investigated kernel methods in which h was fixed across all samples. For example, their h = 4 kernel method used an h of 4 for all samples. The fixed kernel method used here performed better than the best performing kernel method at each sample size in Cope and Kolen.

Discussion and Conclusions

The differences between the methods based on smoothness assumptions and using observed frequencies was smaller for K-S than for MSE, although the

pattern of results were generally the same for both criteria. In fact, for the licensure test for samples of 5000, K-S for the variable kernel method is larger than K-S for the observed frequencies. The smoothness assumptions have less of a positive effect in estimating the cumulative distribution than in estimating the raw score probabilities.

All estimation methods studied here using smoothness assumptions performed far better than using observed frequencies, especially for smaller sample sizes. For example, using the 4-parameter beta binomial method in samples of size 500 yields a lower MSE than using observed frequencies with samples of size 5000 for the ACT Mathematics and Social Science tests, and almost as low a value of MSE using observed frequencies with samples of size 5000 for the licensure test. The differences between the methods based on smoothness assumptions are small compared to the difference between the worst performing method and using observed frequencies.

The results of the simulation study, although limited by the fact that only three population distributions were examined, suggest a preference for the 4-parameter beta binomial method. For the largest sample size the polynomial method performs similarly to the 4-parameter beta binomial method. It has been my experience that both the 4-parameter beta binomial method and the polynomial give good fits to a wide variety of sample distributions.

The polynomial method performed better than the kernel methods except in cases in which the variation of the sample relative frequencies around the true relative frequencies tended to be large, either because of a small sample size and/or a flat distribution. For example, the polynomial method performed worse than the kernel method for the ACT Mathematics data for sample sizes

less than 5000. Figure 3 shows that the ACT Mathematics distribution is relatively flat compared to the other two population distributions.

The poor performance of the polynomial method for the ACT Mathematics test with smaller sample sizes, relative to the kernel methods, may partly be due to the model selection strategy used. The logical flaws of this model selection strategy for choosing an estimate have been mentioned previously.

In practice, a decision of which density estimate to use must be based on the sample data. An estimate of ASE obtained by the bootstrap, jackknife, or cross validation would be useful (although this would require extensive computations for the polynomial method, and to a lesser extent, the 4-parameter beta binomial method), but should probably not be used as the only information on which to choose an estimate. For example, in Figure 2 the approximate cross validation procedure picks h = 2 for fixed kernel method. It is likely that a value of h greater than 2 would be more appropriate here.

For the polynomial and kernel methods it is probably unwise to use an automatic procedure such as those used in the simulation study to choose an estimate. The most practical information to use in deciding on an estimate would be plots of the fitted and raw distributions, chi-square goodness of fit statistics and fitted versus raw sample moments (for the kernel methods). It is suggested that for the polynomial and kernel methods the actual estimate to be used be chosen by looking at the fits for various degrees and window widths and making a judgment based on this information, rather than using automatic procedures as in the study. The results reported in Tables 2 through 4 may either over or underestimate the performance of such subjective procedures for the kernel and polynomial methods depending on the biases of the person choosing the estimate for over or under smoothing and the true distribution.

References

Conover, W. J. (1980). Practical nonparametric statistics. (2nd ed.). New York: John Wiley and Sons.

Cope, R. T., & Kolen, M. J. (1987). A study of methods for estimating distributions of test scores. Paper presented at the 1987 Annual Meeting of the American Educational Research Association.

Darroch, J., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. Annals of Mathematical Statistics, 43, 1470-1480.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia, PA: Society for Industrial and Applied Mathematics (Monograph No. 38).

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of the American Statistical Association, 78, 316-331.

Haberman, S. (1974). Log-linear models for frequency tables with ordered classifications. Biometrics, 30, 589-600.

Johnson, N. L., & Kotz, S. (1970). Continuous univariate distributions-2. New York: John Wiley and Sons.

Kolen, M. J. (1988). Applications of test score distribution estimation/smoothing methods. Paper presented at the 1988 Annual Meeting of the American Educational Research Association, New Orleans (April).

Lord, F. M. (1964). A strong true-score theory, with applications. Educational Testing Service Research Bulletin 64-19. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1965). A strong true-score theory with applications. Psychometrika, 30, 239-270.

Rosenbaum, P. R., & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. British Journal of Mathematical and Statistical Psychology, 40, 43-49.

Tukey, J. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Wong, H. W. (1983). A note on the modified likelihood for density estimation. Journal of the American Statistical Association, 78, 461-463.

TABLE 1

Window Width (h) and Associated Values of Err Chosen Using the Bootstrap and Approximate Cross Validation Estimates of Err for 10 Simulated Samples

	h for Minimum Err		Bootstrap		Approximate Cross Validation	
Sample	h	Err	h	Err	h	Err
1	20	-.026083	20	-.026083	24	-.026082
2	12	-.026076	28	-.026049	36	-.026031
3	36	-.025999	20	-.025976	16	-.025959
4	24	-.025994	8	-.025919	12	-.025960
5	20	-.026057	8	-.026007	12	-.026043
6	24	-.025912	20	-.025912	16	-.025906
7	36	-.026032	24	-.026024	24	-.026024
8	28	-.025999	24	-.025997	20	-.025993
9	24	-.025963	24	-.025963	12	-.025942
10	24	-.025995	12	-.025977	12	-.025977
Mean		-.026011		-.025991		-.025992
s.d.		.0000526		.0000536		.0000531

TABLE 2

Fit of Licensure Test Estimated Densities

Sample Size	Measure of Fit	Unsmoothed Sample Frequencies	Variable Kernel	Fixed Kernel	Polynomial	4-Parameter Beta Binomial
	Bias Squared	.045	2.030	1.583	.783	.769
	Variance	31.699	4.747	4.969	3.852	2.408
500	MSE	31.744	6.778	6.552	4.636	3.178*
	K-S	32.595	28.505	27.835	24.236	21.842*
	(s.e. K-S)	(.505)	(.475)	(.467)	(.517)	(.475)
	Bias Squared	.021	1.527	1.011	.581	.713
	Variance	15.879	2.722	2.894	1.770	1.193
1000	MSE	15.900	4.250	3.905	2.351	1.906*
	K-S	23.519	22.218	20.557	17.331	16.442*
	(s.e. K-S)	(.367)	(.363)	(.348)	(.365)	(.356)
	Bias Squared	.016	1.034	.633	.505	.593
	Variance	8.077	1.553	1.642	1.000	.600
2000	MSE	8.093	2.587	2.275	1.505	1.193*
	K-S	16.296	16.317	14.540	12.875	11.617*
	(s.e. K-S)	(.261)	(.267)	(.254)	(.242)	(.248)
	Bias Squared	.007	.691	.441	.411	.634
	Variance	3.089	.708	.775	.435	.229
5000	MSE	3.096	1.399	1.216	.846*	.863
	K-S	10.234	11.588	10.010	8.614*	8.773
	(s.e. K-S)	(.169)	(.169)	(.161)	(.162)	(.165)

Notes: Values of bias squared, variance, and MSE have been multiplied by 1000000, so they are in terms of frequencies for a sample size of 1000.

Within each sample size the lowest values of MSE and K-S are identified by an '*'.

TABLE 3

Fit of ACT Mathematics Estimated Densities

Sample Size	Measure of Fit	Unsmoothed Sample Frequencies	Variable Kernel	Fixed Kernel	Polynomial	4-Parameter Beta Binomial
	Bias Squared	.114	.632	.925	1.423	.704
	Variance	46.472	5.899	4.911	6.952	3.604
500	MSE	46.586	6.531	5.836	8.375	4.308*
	K-S	33.964	24.923	24.078	28.267	23.718*
	(s.e. K-S)	(.503)	(.486)	(.481)	(.507)	(.490)
	Bias Squared	.048	.571	.798	.890	.624
	Variance	23.715	3.141	2.841	4.085	1.745
1000	MSE	23.763	3.712	3.639	4.975	2.369*
	K-S	24.582	18.409	18.119	21.196	17.463*
	(s.e. K-S)	(.361)	(.345)	(.351)	(.359)	(.352)
	Bias Squared	.0157	.538	.656	.200	.524
	Variance	11.735	1.698	1.690	2.223	.844
2000	MSE	11.752	2.236	2.346	2.423	1.368*
	K-S	16.824	13.174	13.323	14.008	12.476*
	(s.e. K-S)	(.253)	(.243)	(.236)	(.262)	(.246)
	Bias Squared	.007	.416	.355	.093	.438
	Variance	4.689	.762	.916	.666	.319
5000	MSE	4.696	1.178	1.271	.759	.757*
	K-S	10.552	8.989	8.978	8.153*	8.468
	(s.e. K-S)	(.151)	(.147)	(.133)	(.144)	(.138)

Notes: Values of bias squared, variance, and MSE have been multiplied by 1000000, so they are in terms of frequencies for a sample size of 1000.

Within each sample size the lowest values of MSE and K-S are identified by an '*'.

TABLE 4

Fit of ACT Social Studies Estimated Densities

Sample Size	Measure of Fit	Unsmoothed Sample Frequencies	Variable Kernel	Fixed Kernel	Polynomial	4-Parameter Beta Binomial
	Bias Squared	.112	.338	.830	.279	.307
	Variance	36.045	4.609	3.712	4.542	2.189
500	MSE	36.157	4.947	4.542	4.821	2.496*
	K-S	34.001	25.296	25.828	26.237	22.167*
	(s.e. K-S)	(.498)	(.466)	(.455)	(.550)	(.469)
	Bias Squared	.035	.292	.562	.210	.327
	Variance	18.029	2.387	2.198	1.727	1.048
1000	MSE	18.064	2.679	2.759	1.937	1.375*
	K-S	23.831	18.131	19.316	17.121	15.914*
	(s.e. K-S)	(.350)	(.327)	(.310)	(.344)	(.326)
	Bias Squared	.020	.233	.333	.167	.338
	Variance	9.075	1.301	1.276	.908	.510
2000	MSE	9.095	1.534	1.609	1.075	.848*
	K-S	17.017	13.704	14.402	12.551	11.791*
	(s.e. K-S)	(.257)	(.243)	(.237)	(.245)	(.242)
	Bias Squared	.009	.188	.207	.107	.371
	Variance	3.610	.556	.581	.471	.188
5000	MSE	3.619	.744	.788	.578	.559*
	K-S	10.374	9.110	9.404	8.329*	8.427
	(s.e. K-S)	(.160)	(.150)	(.142)	(.157)	(.148)

Notes: Values of bias squared, variance, and MSE have been multiplied by 1000000, so they are in terms of frequencies for a sample size of 1000.

Within each sample size the lowest values of MSE and K-S are identified by an '*'.

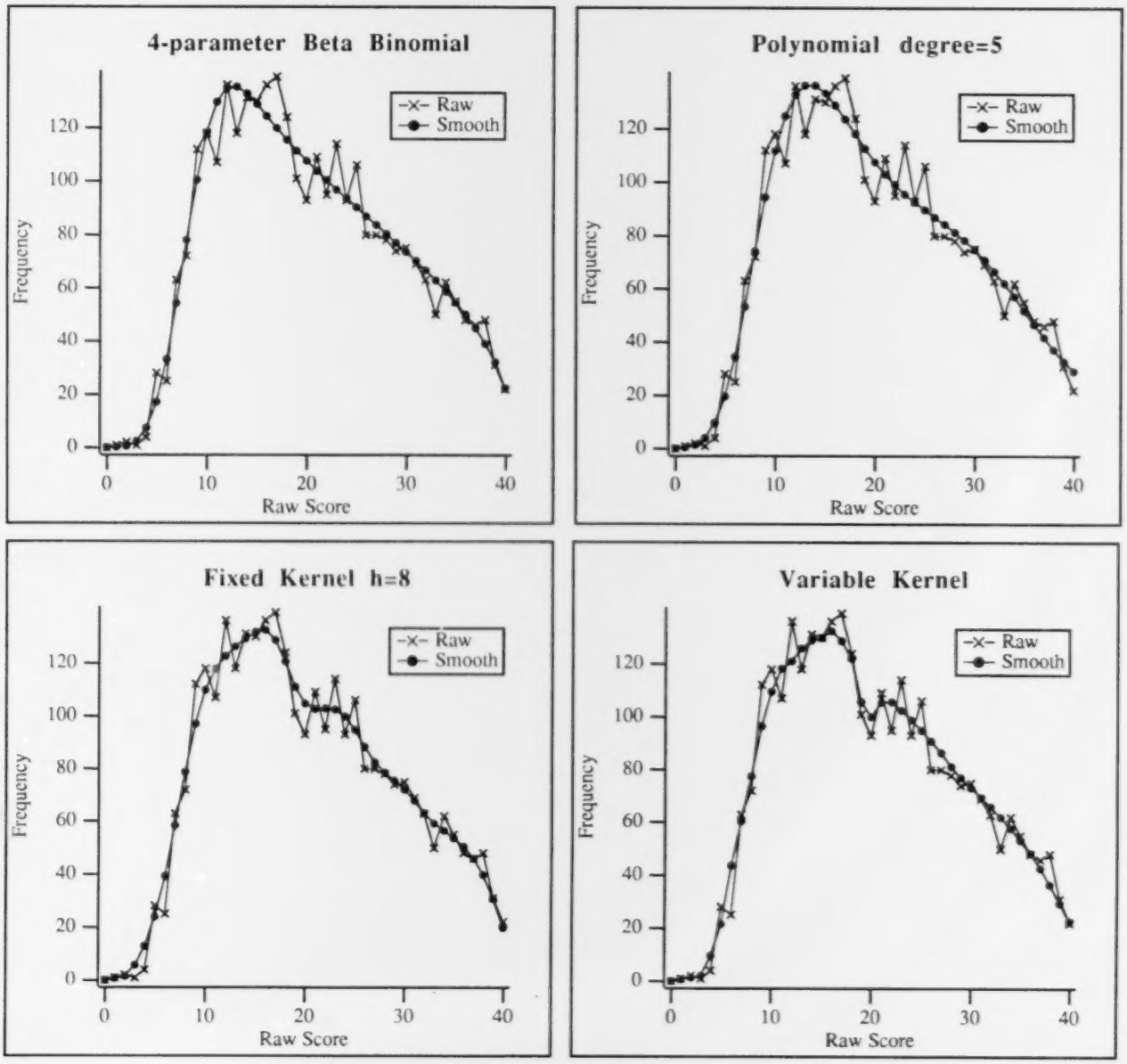

Figure 1. Estimates for ACT Mathematics Test (sample size = 3039)

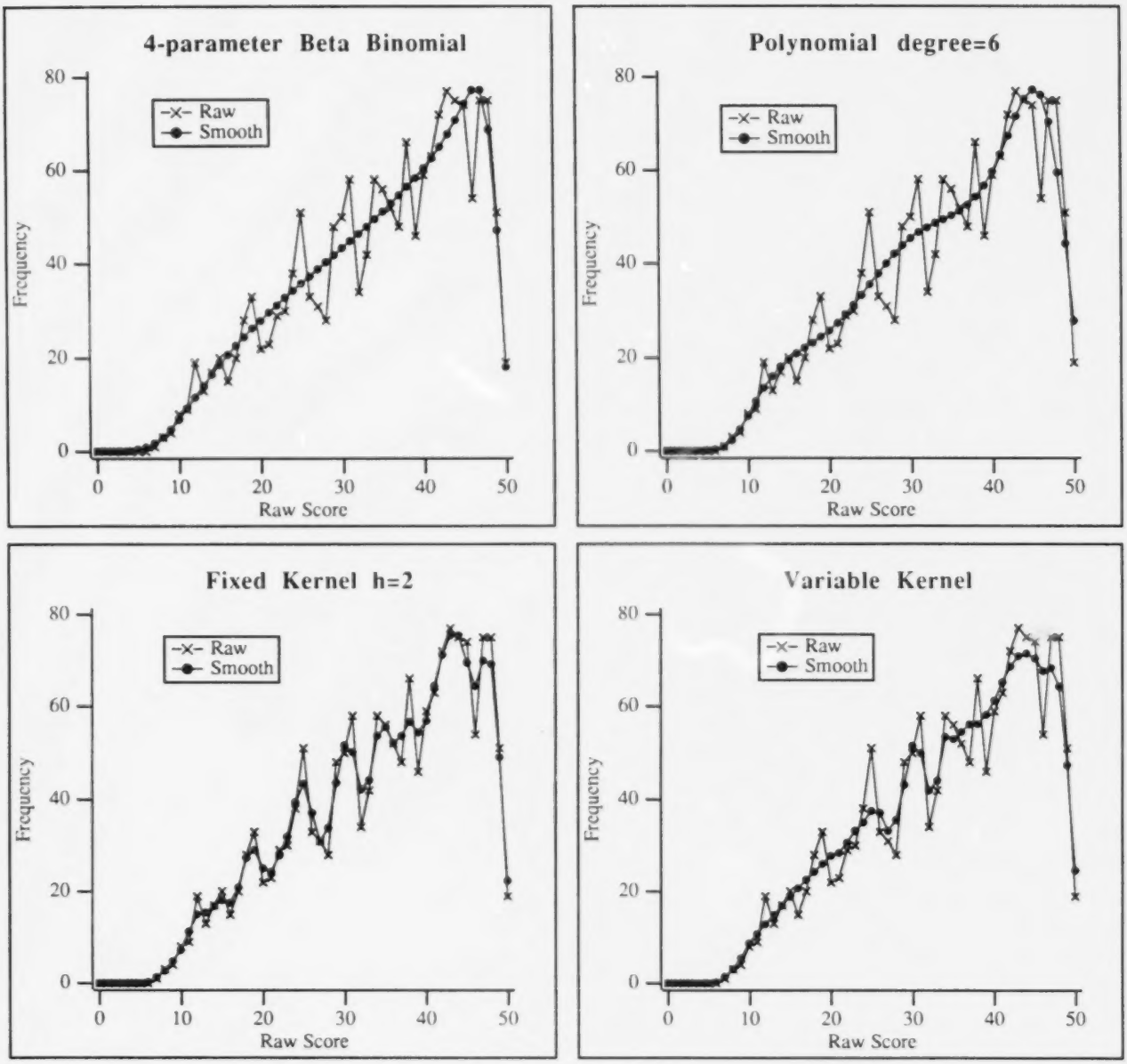

Figure 2. Estimates for P-ACT+ Writing--College-Bound.11th Grade (sample size 1727)

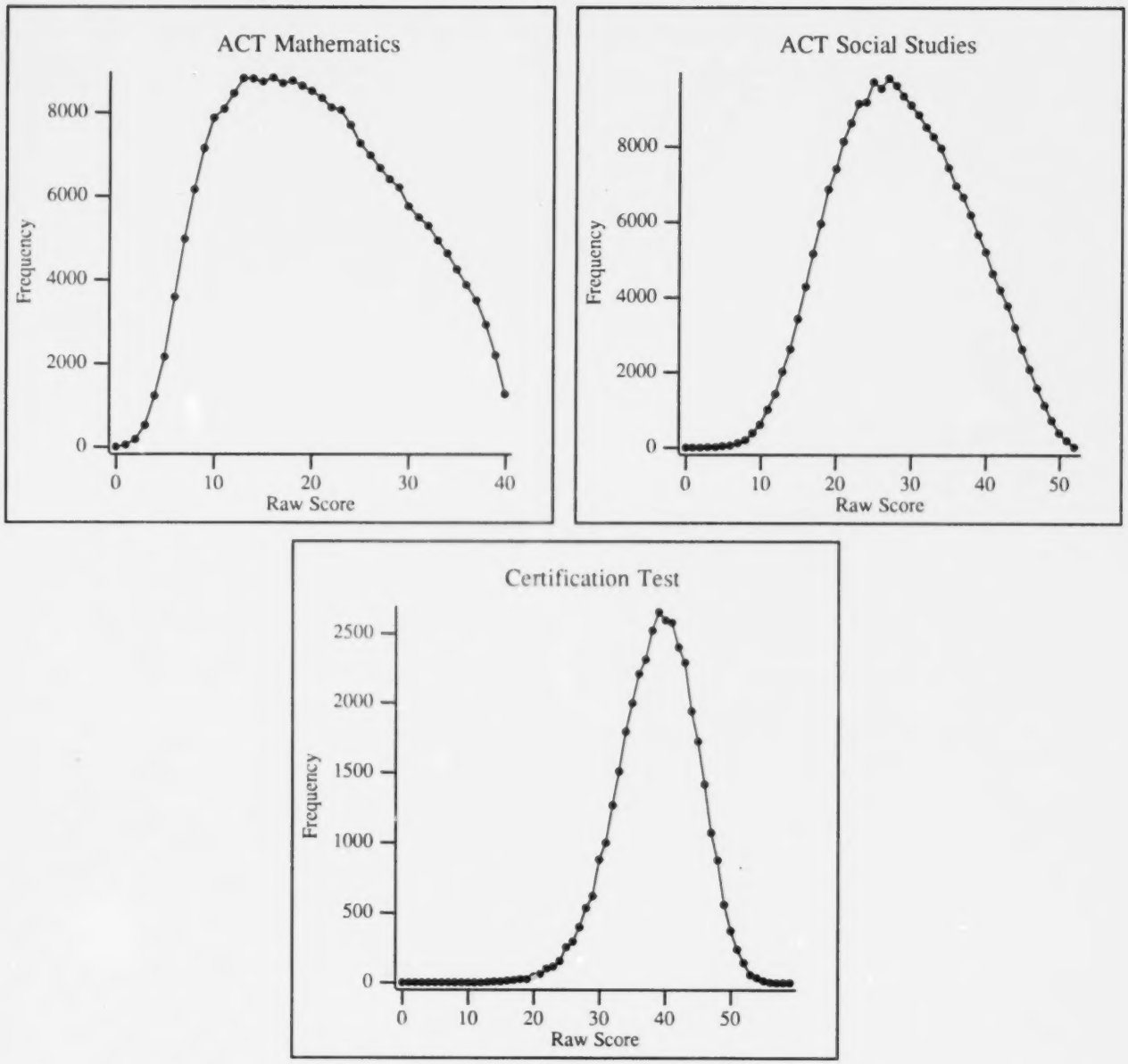

Figure 3. Population Distributions Used in Study

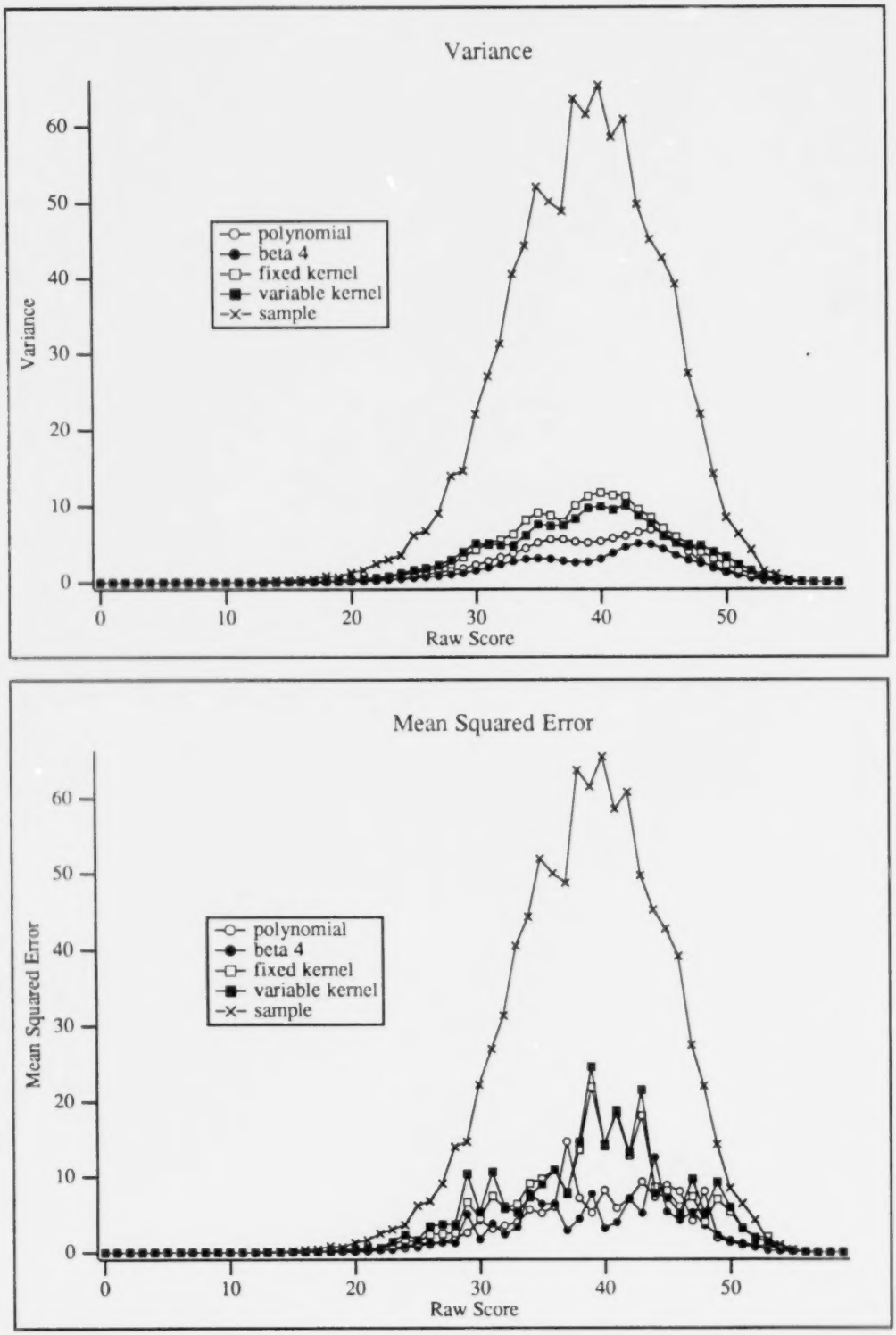

Figure 4. Variance and MSE by Score Point for Licensure Test (N = 1000)

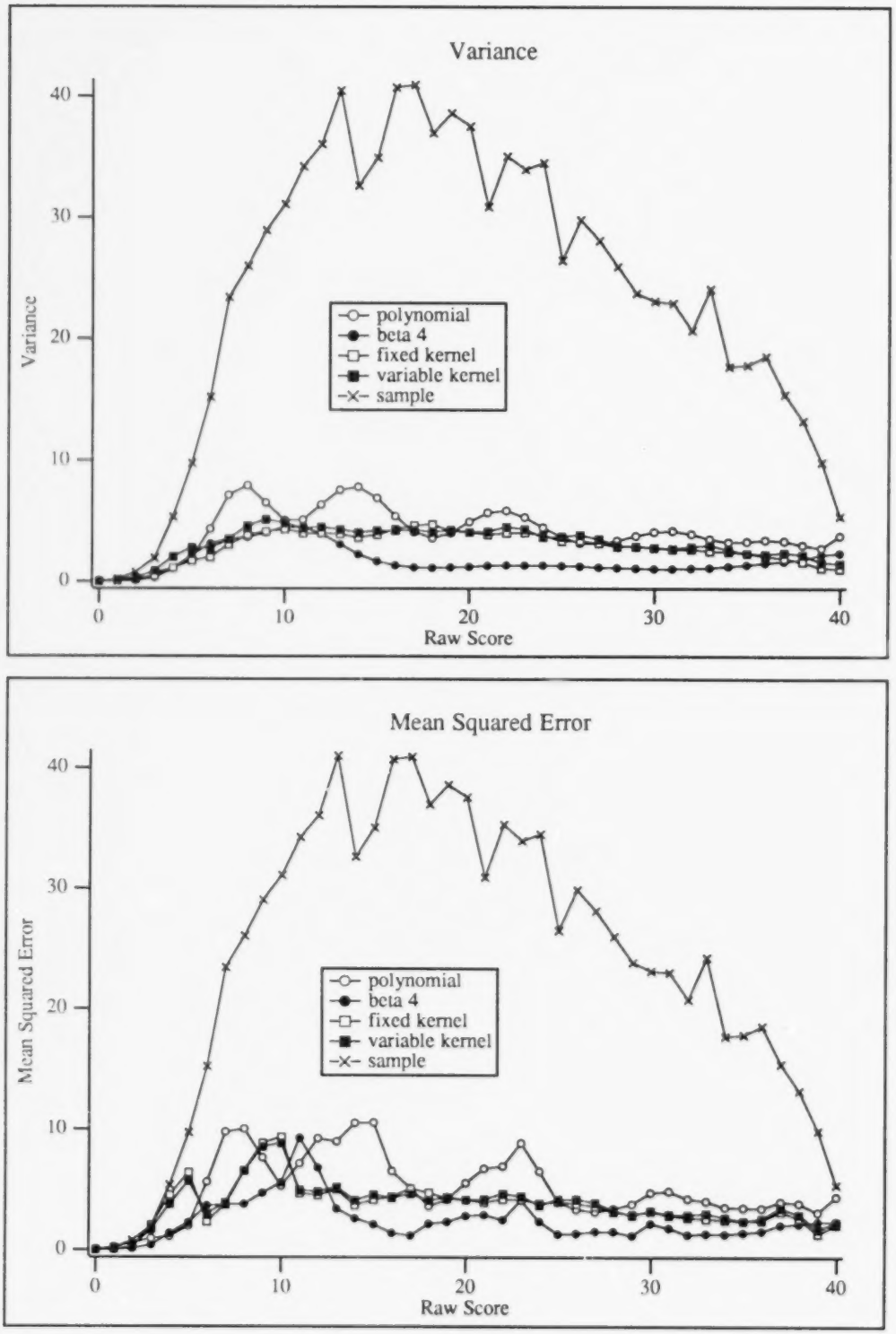

Figure 5. Variance and MSE by Score Point for ACT Mathematics Test (N = 1000)